REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704-0188

1. AGENCY USE ONLY *(Leave blank)*	2. REPORT DATE	3. REPORT TYPE AND DATES COVERED
		FINAL 15 FEB 93 TO 14 JUL 96

4. TITLE AND SUBTITLE
NEURAL NETWORKS FOR LOCALIZED FUNCTION APPROXIMATION

5. FUNDING NUMBERS
F49620-93-1-0150
61102F 2304/HS

6. AUTHOR(S)
H. N. MHASKAR

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)
CALIFORNIA STATE UNIVERSITY
DEPARTMENT OF MATHEMATICS
LOS ANGELES, CA 90032

8. PERFORMING ORGANIZATION
AFOSR-TR-96

0555

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)
AFOSR/NM
110 DUNCAN AVE, SUITE B115
BOLLING AFB DC 20332-8080

10. SPONSORING/MONITORING AGENCY REPORT NUMBER
F49620-93-1-0150

11. SUPPLEMENTARY NOTES

12a. DISTRIBUTION AVAILABILITY STATEMENT
APPROVED FOR PUBLIC RELEASE:
DISTRIBUTION UNLIMITED

12b. DISTRIBUTION CODE

13. ABSTRACT *(Maximum 200 words)*
SEE REPORT

14. SUBJECT TERMS			15. NUMBER OF PAGES
			16. PRICE CODE
17. SECURITY CLASSIFICATION OF REPORT	18. SECURITY CLASSIFICATION OF THIS PAGE	19. SECURITY CLASSIFICATION OF ABSTRACT	20. LIMITATION OF ABSTRACT
UNCLASSIFIED	UNCLASSIFIED	UNCLASSIFIED	SAR

FINAL REPORT

Grant number: F49620-93-1-0150
Project title: Neural networks for localized function approximation
Principal Investigator: H. N. Mhaskar
Department of Mathematics, California State University,
Los Angeles, California 90032.
email: hmhaska@calstatela.edu, **Phone:** 213-343-2157.
Report Period: February 15, 1993- July 15, 1996.

Abstract.
We studied the complexity problem for neural networks used in function approximation; i.e., the problem of estimating the number of neurons needed to provide a given accuracy of approximation for any function, unknown except for a few a priori assumptions. We developed a unified theory, applicable to the traditional neural networks, radial basis function networks, and generalized regularization networks. While our main objective was to provide a solid theoretical foundation for the subject, we have also developed new training paradigms, where no optimization based technique such as back-propagation is required. Thus, the training of our networks is very simple and entirely free of all the traditional shortcomings, such as local minima. Our ideas were tested to develop neural networks for prediction of time series, and beamforming in phased array antennas. In both cases, we obtained spectacular improvements over previously known results. Our work has resulted in 14 publications. In addition, the grant has facilitated the completion of our book on weighted approximation as well as the fulfillment of our obligations as an invited guest editor for a special issue of Advances in Computational Mathematics on Mathematical Aspects of Neural Networks.

1. Introduction. An artificial neural (mapping) network is a mechanism for a highly parallel computation of functions of several real variables. The basic components of a mapping network are called *principal elements* or *neurons*. These are simple processors equipped with a small local memory and are capable of performing certain simple computations such as taking an inner product or evaluating a transfer function (*activation function*). It may take any number of input variables and produce a real output. For example, the action of a typical neuron is to evaluate an expression of the form $\sigma(\mathbf{w}\cdot\mathbf{x}+b)$ where $\mathbf{x}$ denotes the vector of inputs, the *weights* $\mathbf{w}$ and the *thresholds* b are stored in the local memory and σ is the activation function. These are not the only kind of neurons being considered in practice, but these seem to be the most traditional ones. The neurons are organized in *layers*. The *input layer* fans out its input to all the neurons in the first *hidden layer*. The neurons in the hidden layers can be interconnected in any manner whatever. In a *feedforward network*, the outputs of the neurons in any hidden layer is fanned out to the neurons in the next hidden layer. The *output layer* calculates (and outputs) a weighted sum of the outputs of the neurons in the last hidden layer. The network *learns* (*is trained*) by adjusting the weights and thresholds in the various neurons.

It is obvious that a mapping network evaluates a special kind of function. For example, if the network has only one hidden layer, and all the neurons evaluate the same activation function σ, then the output of this network will have the form $\sum_{k=1}^{n} c_k\sigma(\mathbf{w}_k\cdot\mathbf{x}+b_k)$. Many

applications of neural networks, such as guidance and control of an airplane, target classification, analysis of time series, and robotics, involve the approximation of an unknown *target* function. Therefore, the main questions in the theory of mapping networks arise out of the desire to represent an "arbitrary" function at least approximately using networks with one or more hidden layers.

Prior to the start of this project, many authors [22, 25, 27] had studied the *density problem;* i.e., the problem to determine if an "arbitrary" function can be represented within an arbitrarily small margin of tolerance by the outputs of a neural network. However, there was only a scant study [17] concerning the size complexity of the networks required to achieve a prescribed order of approximation. Of course, neural networks were used to solve practical problems involving function approximation in spite of the lack of theory. However, the techniques used are often ad hoc, and require a detailed knowledge of the specific problem being solved.

We observe that although networks which represent boolean functions were studied extensively, approximation of real functions is an entirely different problem with totally different issues of interest. For example, one major problem in the realization of boolean functions is to avoid over-training, rote memorization in the extreme case, as it limits the capacity of the network to generalize. In contrast, a rote memorization is impossible for real functions and the very notion of generalization takes on a new meaning – approximating the target function at points where no training data is available.

2. Detailed description of our work. The objective of the program was to conduct a thorough theoretical investigation of the capabilities of neural networks to approximate functions. In particular, the project aimed at the construction of networks with a minimal number of neurons so as to provide universal approximation of functions when the only a priori knowledge about the target function is that it has a certain number of bounded derivatives. The architecture of the network should be the same for a wide class of functions; only the parameters may depend upon the individual function. Moreover, the approximation should be localized in the sense that if the target function has to be changed on a small part of its domain, then only a few neurons, rather than the whole network, should be retrained.

In general, the target function is an unknown function. However, it is customary to assume that the function belongs to a known function class. For example, the now well known result of Barron [17] shows that if the function is assumed to satisfy certain conditions expressed in terms of its Fourier transform, and each of the neurons evaluates a sigmoidal activation function, then at most $\mathcal{O}(\epsilon^{-2})$ neurons are needed to achieve the order of approximation ϵ. It is sometimes difficult to verify whether the conditions required to apply Barron's theorem are satisfied. It is more customary to assume only that the target function has a certain number of derivatives.

Thus, a common choice of the function class is the Sobolev class $W_{r,s}$, for some integer $r \geq 1$. This class consists of all functions on $[-1,1]^s$ having continuous partial derivatives up to order r. For the sake of clarity of exposition, we limit ourselves to continuous functions, although most of our results are also valid for other L^p classes. In the sequel, we find it convenient to write $1/n$ for ϵ. The symbol $\tilde{E}_{\phi,r,s;n}$ will denote the number of neurons, each evaluating an activation function ϕ, required to yield an approximation order

$1/n$ for *every* function in $W_{r,s}$.

General results in the theory of approximation of functions suggest that $\tilde{E}_{\phi,r,s;n} \geq cn^{r/s}$ for a suitable constant $c > 0$. In fact, the same lower bound is valid for *any* approximation process depending upon n parameters selected in a robust manner. Therefore, we addressed primarily the following questions. (1) What order of approximation can one achieve with an "arbitrary" activation function? (2) Can one achieve the optimal order of approximation for some specific activation function? (3) Are there some limitations on neural networks in terms of the degree or manner of approximation, in spite of their well known universal approximation properties? (4) Is it possible to provide simple training paradigms for the networks which provide the theoretical bounds on approximation? We also investigated a few side issues related to these main problems.

All the networks that we have developed for approximating functions from the Sobolev class share some important features. None of them use any nonlinear optimization. Thus, they are all free of all the pitfalls of such commonly used procedures as backpropagation, e.g. local minima. In fact, our networks define linear operators given by explicit formulas. In particular, their training is especially easy.

In [1], we have proved that it is possible to arrange $\mathcal{O}(n^{s/r})$ neurons, each evaluating a univariate sigmoidal function of order k, where $k \geq 2$ is an integer, in sufficiently many hidden layers so that an arbitrary function in $W_{r,s}$ can be approximated within $1/n$ by the output of the resulting network. The networks in [1] also provided localized approximation in the following sense. The network can be implemented using only a fixed number of neurons, *independent of the desired accuracy.* If the value of the target function is desired at a point $\mathbf{x}$, one trains this network by choosing the training data in a neighborhood of $\mathbf{x}$, using the fast training algorithm described in [1]. The diameter of this neighborhood depends upon the desired accuracy, but it is not required to find the nearest neighbors, or to solve any other optimization problem. From a global perspective, the domain of approximation is divided into small subregions, with diameter depending upon the accuracy desired, and the network consists of subnetworks, each having a small size and each "responsible" for one subregion. If the function has to be modified in some subregion, then only the neurons responsible for this subregion need to be retrained using a very fast process. This aspect was discussed in detail in [3].

During our research, we became aware of the paper [24] of Girosi, Jones, and Poggio. They have introduced (in a somewhat restrictive form) the notion of *generalized translation networks* (GTN's) which can be described mathematically as follows. Let $1 \leq d \leq s$, $N \geq 1$ be integers, and $\phi : \mathbf{R}^d \to \mathbf{R}$. A *generalized translation network* with N neurons evaluates a function of the form $\sum_{k=1}^{N} a_k \phi(A_k(\cdot) + \mathbf{b}_k)$ where the *weights* A_k's are $d \times s$ real matrices, the *thresholds* $\mathbf{b}_k \in \mathbf{R}^d$ and the *coefficients* $a_k \in \mathbf{R}$ $(1 \leq k \leq N)$. The set of all such functions (with a fixed N) will be denoted by $\Pi_{\phi;N,s}$. When $d = 1$, $\Pi_{\phi;N,s}$ is the set of all outputs of a neural network with N neurons, each evaluating the activation function ϕ, and receiving s inputs. When $d = s$, and ϕ is a radially symmetric function, then $\Pi_{\phi;N,s}$ denotes the set of all outputs of a radial basis function network. In [24], Girosi, Jones, and Poggio have pointed out the importance of the study of the more general case considered here. They have demonstrated how generalized translation networks arise naturally in such applications as image processing and graphics, as solutions of certain extremal problems.

In the sequel, the symbol $\tilde{E}_{\phi,r,s;n}$ will denote the number N of neurons required to obtain the degree of approximation $1/n$ for all functions in $W_{r,s}$ by GTN's in $\Pi_{\phi;N,s}$.

In [7], we have studied the complexity problem for the generalized translation networks. The focus of this work was to investigate what activation functions ϕ can give what degree of accuracy – the intention being to compare different activation functions in terms of the number $\tilde{E}_{\phi,r,s;n}$. Among the networks included in this study are the classical neural networks as well as classical radial basis function networks evaluating the Gaussian function, thin plate splines, generalized multiquadrics and other general functions. Our work is the first of its kind in the theory of radial basis function approximation – the degree of approximation is estimated in terms of the number of evaluations of the basis function rather than in terms of a scaling factor. The networks also provide a simultaneous approximation of derivatives of the target function, thus solving a problem that occurs often in control theory (cf. [20, 23]). We have also studied the complexity problem in terms of the number of observations of the function, rather than in terms of the number of neurons involved. The paper [5] contains an announcement of some of these results.

In [8], we have studied a special class of activation functions, satisfying certain smoothness conditions. We have constructed networks that give the optimal order of approximation for Sobolev classes, i.e., for which $\tilde{E}_{\phi,r,s;n} \le cn^{s/r}$. Some of the important examples of the activation functions included among those in [8] are : the squashing function, $(1+e^{-x})^{-1}$, generalized multiquadrics, certain thin plate splines, and the Gaussian function. The weights and thresholds in our networks are all uniformly bounded. The results and ideas in [8] have been applied to the problems of probability density estimation [30] and pattern recognition [18].

In [11], we prove that the coefficients of **any** network using an activation function smoother than the target function must satisfy certain lower bounds; in particular, must become unbounded as the desired accuracy of approximation increases. Our results can be interpreted as a test of the hypothesis about the smoothness of the unknown target function.

Although the activation functions studied in [8] include the most commonly used radial basis functions, they are not of the *pure translation* form $\sum a_k \Phi(||\mathbf{x}-\mathbf{x}_k||)$, where no matrices are involved. In [9], we constructed Gaussian networks of the pure translation form which provide an optimal approximation for functions in the Sobolev class. The centers of this network may be chosen independently of the target function, and arbitrarily close to the origin. An additional novelty of this work is the construction of networks capable of providing approximation on the whole Euclidean space. The networks provide simultaneous approximation of the derivatives and also of the *Fourier transform* of the target function, using information about the function (in the space domain) alone. The proofs of the results in [9] are given in our book [16].

In [2], we consider the problem of localized approximation. We proved that even in the case of the Heavyside activation function, it is not possible to approximate the characteristic function of a two dimensional square by neural networks with one hidden layer consisting of a fixed number of neurons. Thus, the constructions similar to those in [1] cannot be made with networks with one hidden layer. On the positive side, we constructed the Chui-Wang spline wavelets (cf. [21]) using neural networks with multiple

hidden layers and using sigmoidal functions of order at least 2 as activation function. The number of neurons in the approximation of the mother wavelets are independent of the degree of approximation required. This research has motivated further work by Kurkova [26] regarding the nature of functions that can be approximated arbitrarily well by neural networks with a the number of neurons prescribed in advance.

In [10], we strengthened our results of [2]. We proved that with the additional requirement of localization, the number of neurons in a generalized translation network to provide the degree of approximation $1/n$ to all functions in $W_{r,s}$ must be at least $cn^{s/r}\log n$. This is true even when each neuron may evaluate a different activation function, and even when the activation function may depend upon the target function. On the other hand, a neural network using the squashing function as the activation function is capable of giving the same order of approximation, with localization, using at most $n^{s/r+\delta}$ neurons, for any $\delta > 0$.

In [4], we have constructed generalized translation networks to provide *dimension independent* approximation order for the class of functions with summable Fourier coefficients. This generalized the well known result of Barron [17], which was applicable only for sigmoidal activation functions. We also established tight lower bounds for the approximation order. One interesting aspect of this paper is that the conditions on the target function are local. In contrast, the conditions assumed by Barron are in terms of the Fourier transform, which amounts to an assumption on the behavior of the target function on the entire Euclidean space. In fact, we have proved some very general results. If the activation function is an (orthogonal) "mother wavelet", then our results give dimension independent bounds in the case when the wavelet coefficients of the target function are summable.

In our work [13] with our student N. Hahm, we study the problem of system identification. As pointed out by Sandberg [28], this problem can be thought of as the problem of approximating a functional defined on a function space, typically some L^p space, rather than functions defined on a Euclidean space. We have constructed generalized translation networks to uniformly approximate a class of nonlinear, continuous functionals defined on $L^p([-1,1]^s)$ for integer $s \geq 1$, $1 \leq p < \infty$ or $C([-1,1]^s)$. We obtain lower bounds on the possible order of approximation for such functionals in terms of *any* approximation process depending continuously upon a given number of parameters. Our networks almost achieve this order of approximation in terms of the number of parameters (neurons) involved in the network. The training is simple and noniterative; in particular, we avoid any optimization such as that involved in the usual back-propagation. An announcement of these results appears in [14].

The paper [6] is an invited survey paper, reviewing some of our work until the time when it was written.

In the joint work [12] with our undergraduate student, L. Khachikyan, we applied some of the ideas in [1] to the problem of constructing general algorithms for the prediction of time series. Superficially similar to the CART algorithm [19], our algorithm involves a very simple training method, based on adaptive approximation, that does not involve *any* optimization. For the flour data [29], our results are 30 to 40 times better than previously known results.

During the summer of 1996, we collaborated with Dr. H. Southall at the Hanscom Air Force Base on the problem of beam steering using phased array antennas. Some of our ideas were used to construct neural networks which do not utilize any nonlinear optimization. On all the data sets investigated by Dr. Southall's group using traditional method of training radial basis function networks, our new methods gave a dramatic improvement.

3. Conclusions. We have conducted a thorough investigation of the complexity problem in the theory of neural networks; i.e., to determine the number of neurons required to give a prescribed degree of approximation to an unknown target function, which satisfies a minimal a priori assumption that it has a certain number of bounded derivatives. The problem was not studied before, and in general, very little was known at the beginning of our work. Our ideas have lead us to new paradigms for training neural networks. Instead of attempting to find the *best* fit for the data, we find a *good* fit. This is done so as to ensure a desired accuracy, and without using **any** nonlinear optimization. On the practical, diverse, and difficult problems of predicting an economic market, as well as beam steering using phased array antennas, our new paradigms provide a substantial improvement over previously known results.

Publications supported by the grant.

1. Approximation properties of a multilayerd feedforward artificial neural network; Advances in Computational Mathematics, **1** (1993), 61-80.
2. Neural networks for localized approximation; Mathematics of Computation, **63** (1994), 607-623 (With C. K. Chui and X. Li).
3. Neural networks for localized approximation of real functions; In "Neural Networks for Signal Processing III", (C.A. Kamm et. al. eds.), IEEE, New York, 1993, 190-196.
4. Dimension-independent bounds on approximation by neural networks; IBM J. of Research and Development,**38** (1994), 277-284 (With C. A. Micchelli).
5. How to choose an activation function; in "Neural Information Processing Systems, 6", (J. D. Cowan, G. Tesauro, J. Alspector Eds.), Morgan Kaufmann Publishers, San Fransisco, 1993, pp. 319-326 (With C. A. Micchelli).
6. Approximation of real functions using neural networks; in "Proc. of Int. Conf. on Computational Mathematics, New Delhi, India, 1993", (C. A. Micchelli ed.), World Scientific, 1994, pp. 267-278.
7. Degree of approximation by neural and translation networks with a single hidden layer, Advances in Applied Mathematics, **16** (1995), 151-183. (With C. A. Micchelli).
8. Neural networks for optimal approximation of smooth and analytic functions; Neural Computation, **8** (1996), 164- 177.
9. Versatile Gaussian networks; Proceedings of IEEE Workshop on Nonlinear Image and Signal Processing, (I. Pitas Editor), Halkidiki, Greece, June, 1995, IEEE, pp.70-73.
10. Limitations of the approximation capabilities of a neural network with a single hidden layer, Advances in Computational Mathematics, **5** (1996), 233-243. (With C. K. Chui and X. Li).
11. On smooth activation functions; Accepted for publication in the Annals of Mathematics in Artificial Intelligence.

12. Neural networks for function approximation; in "Neural networks for signal processing, V", (F. Girosi, J. Makhoul, E. Manolakos, E. Wilson Eds.), IEEE, New York, 1995, pp.21-29. (With L. Khachikyan).
13. Neural networks for functional approximation and system identification; Accepted for publication in Neural Computation. (With N. Hahm)
14. System identification using neural networks; in "Neural networks for signal processing, VI", (S. Usui, Y. Tohkura, S. Katagiri, and E. Wilson Eds.), pp. 82–88, IEEE, New York, 1996, (With N. Hahm).
15. Mathematical Aspects of Neural Networks, Special Issue of Advances in Computational Mathematics, **5** (1996) (Editor with C. A. Micchelli).
16. "An introduction to the theory of weighted polynomial approximation", World Scientific, Singapore, To appear.

Other cited references.

17. A. R. BARRON, *Universal approximation bounds for superposition of a sigmoidal function,* IEEE Trans. Information Theory, **39** (1993), 930-945.
18. K. L. BLACKMORE, R. C. WILLIAMSON, AND I. M. Y. MAREELS, *Decision region approximation by polynomials or neural networks,* To appear in IEEE Trans. Neural Networks.
19. L. BREIMAN, J. H. FRIEDMAN, R. A. OLSEN, AND C. J. STONE, "Classification and regression trees", Wadsworth and Brooks, Pacific Grove, California, 1984.
20. P. CARDALIAGUET AND G. EUVRARD, *Approximation of a function and its derivative with a neural network,* Neural Networks, **5** (1992), 207-220.
21. C. K. CHUI AND J. Z. WANG, *On compactly supported spline wavelets and a duality principle,* Trans. Amer. Math Soc. **330** (1992), 903–916.
22. G. CYBENKO, *Approximation by superposition of sigmoidal functions,* Mathematics of Control, Signal and Systems, **2** (1989), 303-314.
23. A. R. GALLANT AND H. WHITE, *On learning the derivatives of an unknown mapping with multilayer feedforward networks,* Neural Networks, **5** (1992), 129-138.
24. F. GIROSI, M. JONES AND T. POGGIO, *Regularization theory and neural networks architectures,* Neural Computation, **7** (1995), 219-269.
25. K. HORNIK, M. STINCHCOMBE AND H. WHITE, *Multilayer feedforward networks are universal approximators,* Neural Networks, **2** (1989), 359-366.
26. V. KURKOVA, *Approximation of functions by perceptron networks with bounded number of hidden units,* Neural Networks, **8** (1995), 745–750.
27. J. PARK AND I. W. SANDBERG, *Universal approximation using radial basis function networks,* Neural Computation, **3** (1991), 246-257.
28. I. W. SANDBERG, *Approximation theorems for discrete time systems,* IEEE Trans. Circuits Syst., **38** (1991), 564- 566.
29. G. C. TIAO AND R. S. TSAY, *Model specification in multivariate time series,* J. R. Statist. Soc., **51** (1989), 157–213.
30. A. ZEEVI, R. MEIR, AND V. MAIOROV, *Approximation and estimation bounds for nonlinear regression using mixtures of experts,* To appear in Neural Computation.